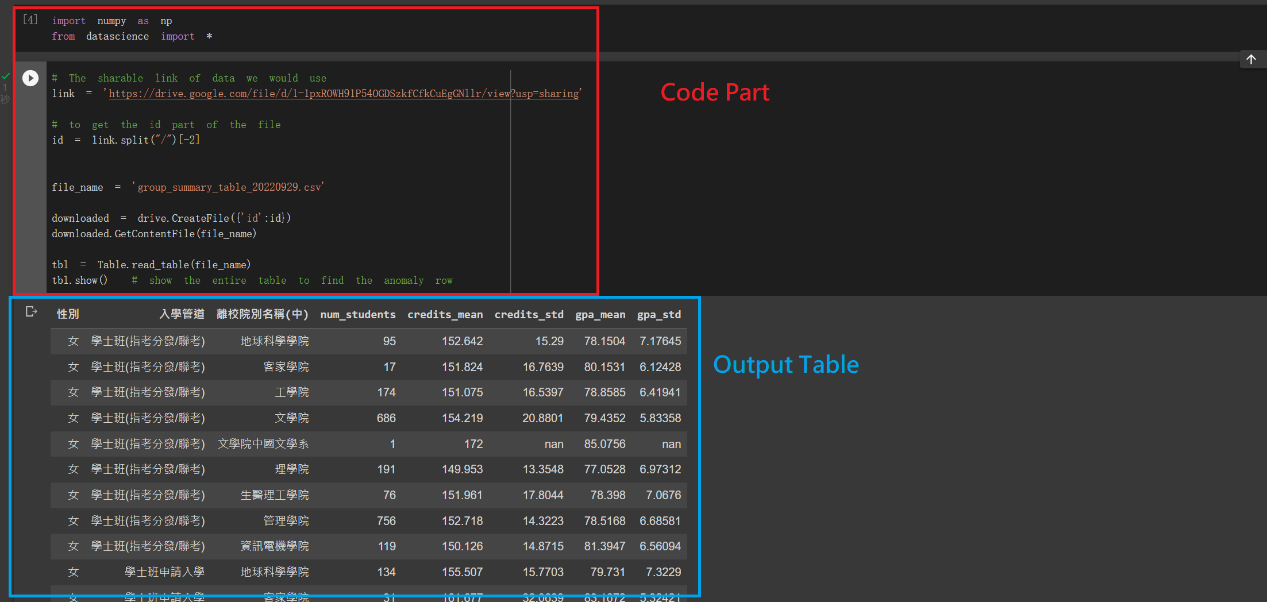
**NCU Introduction to Data Science Fall 2022 – HW1**

NCU MIS 3A 109403019 鄒翔宇

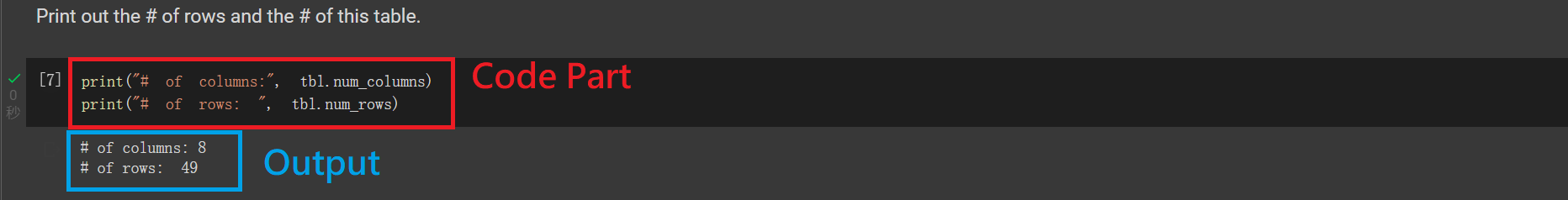
1. **Pleas write code in Colab to read the table and report its size (# of columns and rows)**

First, I use *PyDrive* to load the data in CSV file. In order to use tables, I also import all of the module called *datascience.* Then, I use the Table method *read\_table* to read our imported CSV file and use *show()* method to print out the entire table to find the anomaly row.



Second, I use the *num\_rows* and *num\_columns* methods in Table object type to print out its size.

We can see that the **# of columns is 8** and the **# of rows is 49**.



1. **Which row looks like an anomaly and should be removed? Why?**

I want to find those rows whose values in column *num\_students* is pretty low because they would lead their own mean value and the standard deviation value to be meaningless when comparing with other colleges.

* 1. Please explain.

Ans: Because the number of students is 1, it would cause denominator become 0 when we calculate the standard deviation.

* 1. Please write code in Colab to remove it and return an updated table.

Ans:

1. **Are there more girls than boys?**
   1. Please use `group` to do it.
   2. What % of students are female?
2. **Which college has the most graduates?**
   1. Please use `group` to do it.
   2. Pleas plot a bar chart to show the distribution of # of graduates by college and sort it in descending order.
3. **Which college has the most unbalance 入學管道? In other word , we say it’s balanced if each 管道 has similar number of graduates.**
   1. Please use `group` to show a pivot table, where 入學管道 and college in vertical and horizon directions, respectively, and each cell shows total number of students.
   2. Please plot a bar chart to show the distribution of # of graduates by college and 入學管道.